1 Background

This chapter reviews one of the most widespread uses of wavelets in statistics: nonlinear nonparametric regression. Simple insight into the utility and power of the wavelet approach comes from a discrete spectral analysis point of view. Some of the ideas here are the same as some introduced in Chapter 1, but the different viewpoint is intended to give additional insights.

1.1 Spectral Analysis

Spectral methods for curve estimation are based on "spectral decompositions" of vectors in $\Re^n$. These involve a "transformation", i.e. a change to a different basis of $\Re^n$. Such transformations can provide very powerful and useful new ways to work with given sets of vectors. For example, vectors that are a sum of a few sinusoids, are effectively summarized (and thus studied in many ways) by changing to the Fourier orthonormal basis. Wavelets are simply viewed as another such orthonormal basis, whose good properties are shown in this chapter.

The mathematics of discrete spectral analysis starts with $\{\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_n\}$, an arbitrary orthonormal basis of $\Re^n$, i.e.

$$\boldsymbol{\psi}_i^t \boldsymbol{\psi}_{i'} = \begin{cases} 0 & \text{if } i \neq i' \\ 1 & \text{if } i = i', \end{cases} \tag{1}$$

The spectral representation of a vector $\boldsymbol{\mu} \in \Re^n$ is defined as

$$\mu_i = \sum_{i'=1}^{n} \theta_{i'} \psi_{i',i}, \ i = 1, ..., n, \tag{2}$$

where $\psi_{i',i}$ is the i-th entry of $\boldsymbol{\psi}_{i'}$, and where the coefficients $\theta_{i'}$ are given by the inner products

$$\theta_{i'} = \boldsymbol{\mu}^t \boldsymbol{\psi}_{i'}, \tag{3}$$

which give the lengths of the projection of $\boldsymbol{\mu}$ in the direction of $\boldsymbol{\psi}_{i'}$. The coefficients can be collected into a vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^t$ which is called the "transform" of $\boldsymbol{\mu}$. Useful intuition comes from thinking of θ as a "rotation in $\Re^n$" (only an approximation, since other transformations such as mirror images are also orthonormal).

The particular rotation of this type called the "discrete Fourier Transform" has been a workhorse tool, especially in fields where the study of periodicities and frequencies are important, perhaps most noticeably in electrical engineering. This is because the Fourier rotation of vectors is particularly adept at revealing their frequency structure. In statistics, the best known use of this rotation of data vectors is the Fourier analysis of time series. See Bloomfield (1976) for a very readable account.

A potentially confusing aspect of spectral analysis is that it can often be viewed both "discretely", in terms of ordinary vectors in $\Re^n$, and "continuously"

where the "vectors" are functions in a suitable Hilbert space. The two are typically related to each other because summation (the operation underlying the discrete inner product) is related to integration (the operation underlying the continuous inner product) through Riemann integration. However, orthonormality and the properties following from it are exact in both contexts. For example, in the Fourier basis, the familiar sin and cos waves are orthogonal as functions (with respect to the integral inner product), but also as vectors when evaluated at appropriate equally spaced grids (with respect to summation). The distinctions between these parallel theories is often blurred, because of the approximation. Only the discrete theory is studied in this chapter, with occasional reference made to the continuous theory where it adds insight.

1.2 Nonlinear nonparametric regression

An important statistical application of spectral analysis is to nonparametric regression. Here one models a data vector as

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\mu}$ is some "smooth" underlying mean (i.e. "signal") vector, and $\boldsymbol{\varepsilon}$ is a random "noise" vector (assumed to have mean $\mathbf{0}$, which results in $E(\mathbf{Y}) = \boldsymbol{\mu}$).

Orthogonal transformation allows recasting the problem of using $\mathbf{Y}$ to estimate $\boldsymbol{\mu}$, into that of using the transform of $\mathbf{Y}$ to estimate the transform of $\boldsymbol{\mu}$, which can often be much simpler. Let $\widetilde{\boldsymbol{\theta}} = \left(\widetilde{\theta}_1, ..., \widetilde{\theta}_n\right)^t$ denote the transformation of the data, where the "empirical coefficients" are

$$\widetilde{\theta}_{i'} = \mathbf{Y}^t \boldsymbol{\psi}_{i'}, \tag{4}$$

which are unbiased estimates of the $\theta_{i'}$, the transform coefficients of $\boldsymbol{\mu}$ from (2). A further important property of the rotation aspect of an orthonormal transformation is that when the errors are independent and homoscedastic the "spherical" covariance structure of $\widetilde{\boldsymbol{\theta}}$ is the same as that of $\mathbf{Y}$.

For a good choice of basis, most of the "power of $\boldsymbol{\mu}$" (a useful concept from signal processing), which is conveniently quantified as a sum of squares,

$$P_{\boldsymbol{\mu}} = \sum_{i=1}^{n} {\mu_i}^2 = \sum_{i=1}^{n} {\theta_i}^2$$

(where the last equality follows by the Parseval identity), will be "contained in a few" of the θ_i. In this case, a reasonable reconstruction of the "signal" $\boldsymbol{\mu}$ can be obtained from the data $\mathbf{Y}$ by inverting the transform, but using only the "important" coefficients:

$$\widehat{\mu}_i = \sum_{i' \in S} \widetilde{\theta}_{i'} \psi_{i',i}, \ \ i = 1, ..., n, \tag{5}$$

(sometimes grouped into the vector $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, ..., \widehat{\mu}_n)^t$) where S is some set of "high power coefficients". If the set S is small, but at the same time the

restricted power $\sum_{i' \in S} \theta_{i'}^2$ is a large fraction of $P_{\boldsymbol{\mu}}$, the estimator will be very effective, since most of the power (i.e. component of the sum of squares) of the noise will be contained in terms that do not appear in $\widehat{\boldsymbol{\mu}}$, which will thus be eliminated.

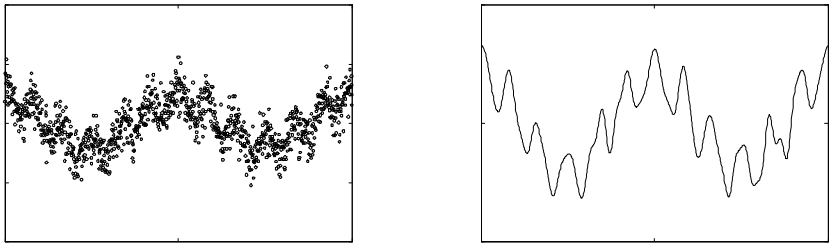

FIGURE 1. Left panel: $n = 1024$ simulated data points, using Gaussian white noise. Right Panel: Fourier basis nonparametric regression estimate, using hard thresholding.

This principle is demonstrated in Figure 1. The left panel shows a simulated data vector $\mathbf{Y}$, which was generated by adding homoscedastic Gaussian noise to a smooth underlying mean vector $\boldsymbol{\mu}$. The right panel shows a signal recovery (i.e. a nonparametric regression estimate), $\widehat{\boldsymbol{\mu}}$, using the Fourier basis, where the set S of coefficients was chosen using the "Universal Hard Thresholding" method discussed in Section 3 of Chapter 1. While the data vector $\mathbf{Y}$ only suggests a single periodic component, note that $\widehat{\boldsymbol{\mu}}$ suggests the underlying mean curve has two periodic components, which is in fact the actual structure of this underlying $\boldsymbol{\mu}$. This illustrates the potential power of the idea of rotating the data vector in a carefully chosen direction, and then using just a few coefficients to dampen noise in signal recovery, i.e. nonparametric regression.

This noise damping effect has a simple, useful quantification when the errors are uncorrelated and homoscedastic. In that case the average variance is:

$$n^{-1} \sum_{i=1}^{n} var\left(\widehat{\mu}_i\right) = \frac{\sigma^2}{n} \#\left(S\right), \tag{6}$$

which is smaller when S has fewer members. On the other hand smaller S means more average squared bias:

$$n^{-1} \sum_{i=1}^{n} \left(E\widehat{\mu}_i - \mu_i\right)^2 = n^{-1} \sum_{i' \notin S} {\theta_{i'}}^2 = n^{-1} \left(P_{\boldsymbol{\mu}} - \sum_{i' \in S} {\theta_{i'}}^2 \right). \tag{7}$$

This shows that the size of S works as a "smoothing parameter" in terms of controlling the usual trade off between variance (which quantitates "wiggliness") and squared bias (which quantitates "goodness of fit"). Also, the spectral type estimator $\widehat{\boldsymbol{\mu}}$ will be most effective when there is a small set S which contains most of the power of the signal, and when that set can be approximately identified. I.e. the potential good performance of $\widehat{\boldsymbol{\mu}}$ is linked to concepts of "signal

compression". An example of this is when the Fourier Transform is applied with a smooth and periodic underlying signal $\boldsymbol{\mu}$, and S is a set of low frequency terms, as shown in Figure 1. However, when the underlying regression function $\boldsymbol{\mu}$ and the basis are such that "power of the signal is spread across the spectrum" (e.g. the Fourier transform of a rough, aperiodic signal), meaning most of $P_{\boldsymbol{\mu}}$ is shared by many of the θ_i, there will not be a good choice of S, and this type of estimator will not perform well. Figure 2 shows an extreme example of this.

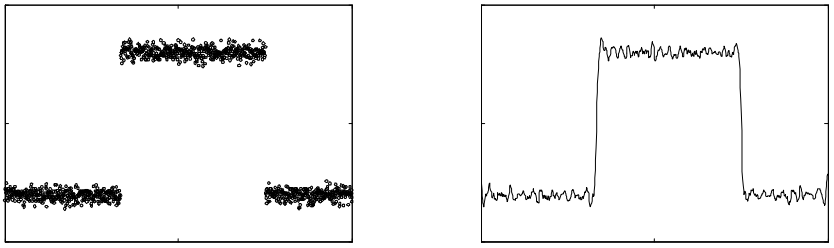

FIGURE 2. Left panel: $n = 1024$ simulated data points, using Gaussian white noise. Right Panel: Fourier basis nonparametric regression estimate, using hard thresholding.

The data in the leftt panel of Figure 2 are again simulated with independent Gaussian noise. It is visually clear that the underlying signal $\boldsymbol{\mu}$ has two large jumps. The Fourier basis does a poor job of compressing such a signal (i.e. the power of the signal is spread all across the spectrum), which has a serious negative impact on the nonparametric regression estimate $\widehat{\boldsymbol{\mu}}$, shown in the right panel. Note that at many locations the wiggliness of $\widehat{\boldsymbol{\mu}}$ is visually almost as large as the range of the data. This is especially true near the jump points, where "ringing" effects, i.e. Gibb's phenomena, are quite noticeable. Furthermore because the Fourier basis is smooth, the crisp jumps become smoothed. Other choices of the set S, result in either increased ringing, or else increased rounding of the jumps. The choice of S used here is (as in Figure 1) the Universal Hard Thresholding method discussed in Section 3 of Chapter 1.

One may wonder why the estimator $\widehat{\boldsymbol{\mu}}$ is called "nonlinear", since the representations (5) and (4) suggest that it is a linear function of the data vector $\mathbf{Y}$. The distinction is in the choice of the set S. When S is chosen independently of the data, e.g. when using the Fourier basis with just a set of low frequency terms, then $\widehat{\boldsymbol{\mu}}$ is a linear estimator. However, when S is chosen as some function of the data $\mathbf{Y}$, e.g. according to various types of thresholding rules as in Section 3 of Chapter 1, then $\widehat{\boldsymbol{\mu}}$ is (perhaps mildly) nonlinear.

2 The Wavelet Bases

The excitement that has surrounded the development of wavelet bases comes from their surprisingly good signal compression of a wide new range of signal types.

2.1 Wavelet Fundamentals

The Fourier basis is effective when signals are "spatially homogeneous", meaning the amount of smoothness is roughly the same in different locations. A set of bases which can be effective when the smoothness is not homogenous are the wavelet bases. Some wavelet bases are good at compressing smooth signals, as well as those that are "somewhat unsmooth in some locations", such as the step function underlying Figure 2. This is accomplished by having basis vectors which are smooth, but allow "localization in time", as well as in frequency.

As with Fourier theory, wavelets have closely parallel discrete and continuous theories, which are connected by Riemann summation. Again, the discrete case is treated here, although sometimes it is useful to think in terms of continuous functions evaluated at $x_1, ..., x_n$, equally spaced on the unit interval, e. g. $x_i = i/n$. The presentation is simplest when n is a power of 2, so that will be assumed throughout this section.

A good starting point for understanding the special structure of wavelet bases is the Haar basis. Some of these basis functions are shown in Figure 4 of Chapter 1. A way of organizing these is shown here in Figure 3. This basis provides coefficients θ_i, conveniently reindexed as $\theta_{j,k}$, which give both a "scale" (this is wavelet terminology that should be viewed as a synonym for "frequency" by those familiar with Fourier analysis) and a "location" decomposition of $\boldsymbol{\mu}$. A convenient index for scale is $j = 0, ..., \log_2(n/2)$. At scale j, basis vectors essentially consist of disjoint step functions (thinking continuously for the moment), indexed by $k = 0, ..., 2^j - 1$, whose intervals of support are the dyadic intervals $(k2^{-j}, (k+1)2^{-j}]$. Orthogonality with coarser scale basis vectors comes from making the step function assume positive and negative values that are equal in magnitude, on the two halves of the support. More precisely (and now thinking discretely), using the vector notation

$$\mathbf{1}(n) = (1, ..., 1)^t, \ \mathbf{0}(n) = (0, ..., 0)^t \text{ and } \mathbf{0}(0) = \{\},$$

define

$$\boldsymbol{\psi}_{j,k} = \left(\frac{2^j}{n}\right)^{1/2} \begin{pmatrix} \mathbf{0}\left(\frac{kn}{2^j}\right) \\ \mathbf{1}\left(\frac{n}{2^{j+1}}\right) \\ -\mathbf{1}\left(\frac{n}{2^{j+1}}\right) \\ \mathbf{0}\left(n - \frac{(k+1)n}{2^j}\right) \end{pmatrix}, \tag{8}$$

for $j = 0, ..., \log_2(n) - 1$ and $k = 0, ..., 2^j - 1$. The factor of $\left(\frac{2^j}{n}\right)^{1/2}$ makes the lengths one. Basis vectors with the same scale j are orthogonal because the nonzero entries are disjoint. Orthogonality across frequencies follows either from this disjointness, or from inner products resulting in a subvector of the form $\begin{pmatrix} \mathbf{1} \\ -\mathbf{1} \end{pmatrix}$ being multiplied by a constant vector. These $n - 1$ vectors, together with $\boldsymbol{\varphi}_{0,0} = n^{-1/2}\mathbf{1}(n)$ form an orthonormal basis of $\Re^n$, which is visually represented in Figure 3.

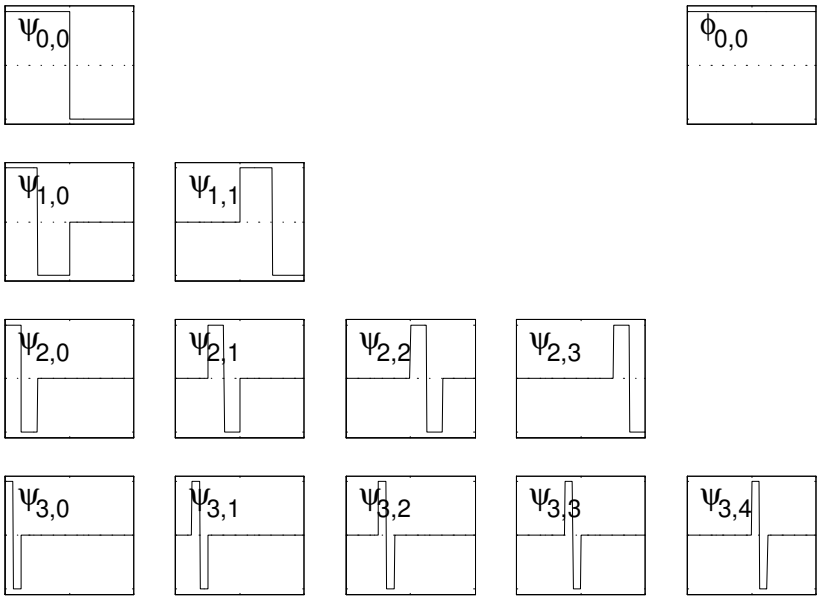

FIGURE 3: The Haar basis, with "father vector" $\varphi_{0,0}$ and "mother vectors" $\psi_{j,k}$. The row for scale $j = 3$ is to the right. The pattern continues below the bottom for scales $j = 4, 5, ...$

The Fast Fourier Transform was considered to be a major computational breakthrough, since it allows calculation of the n Fourier coefficients with only $O(n \log n)$ computations, instead of the $O(n^2)$ operations needed for the obvious naive computation using (4). The key is a very clever reorganization of the computation, which makes heavy use of special properties of trigonometric functions. The Haar coefficients $\theta_{j,k}$ can also be calculated very efficiently by clever reorganization, which turns out to be both simpler and even faster than the FFT.

Insight into the fast algorithm comes from the relationship (Section 2.2 of Chapter 1 gives another way of looking at this) between the "mother basis vectors" $\boldsymbol{\psi}_{j,k}$ from (8), and the "father basis vectors"

$$\boldsymbol{\varphi}_{j,k} = \left(\frac{2^j}{n}\right)^{1/2} \begin{pmatrix} \mathbf{0}\left(\frac{kn}{2^j}\right) \\ \mathbf{1}\left(\frac{n}{2^j}\right) \\ \mathbf{0}\left(n - \frac{(k+1)n}{2^j}\right) \end{pmatrix},$$

for $j = 0, ..., \log_2(n)$ and $k = 0, ..., 2^j - 1$. Note that for any j_0, $\boldsymbol{\varphi}_{j_0,0}, ..., \boldsymbol{\varphi}_{j_0,2^{j_0}-1}$ and $\boldsymbol{\varphi}_{0,0}, \boldsymbol{\psi}_{0,0}, \boldsymbol{\psi}_{1,0}, ..., \boldsymbol{\psi}_{j_0-1,2^{j_0-1}}$ span the same subspace (essentially step functions over intervals of length $\frac{1}{2^{j_0}}$) of $\Re^n$. In Section 2.2 of Chapter 1, this subspace is denoted as V_{j_0}. Hence the "tip of the pyramid" shown in Figure 3 (down to any given row) can be replaced by a row of father basis vectors, to

yield another orthogonal basis. The wavelets also have a "magnification property" in that (modulo location) higher frequency basis vectors $\boldsymbol{\varphi}_{j,k+1}$ ($\boldsymbol{\psi}_{j,k+1}$) are simple horizontal rescalings of $\boldsymbol{\varphi}_{j,k}$ ($\boldsymbol{\psi}_{j,k}$ respectively), by a factor of 2. This is why the index j is said to index "scale", even though "frequency" is also a useful interpretation.

The crucial relationship between these is that the mother and father basis vectors at each frequency are a simple linear function of father vectors at the next higher frequency:

$$\begin{aligned} \boldsymbol{\psi}_{j,k} &= \tfrac{1}{\sqrt{2}} \left(\boldsymbol{\varphi}_{j+1,2k} - \boldsymbol{\varphi}_{j+1,2k+1} \right) \\ \boldsymbol{\varphi}_{j,k} &= \tfrac{1}{\sqrt{2}} \left(\boldsymbol{\varphi}_{j+1,2k} + \boldsymbol{\varphi}_{j+1,2k+1} \right). \end{aligned} \tag{9}$$

Since most of the wavelet coefficients are of the inner product form $\theta_{j,k} = \boldsymbol{\mu}^t \boldsymbol{\psi}_{j,k}$, they can be quickly obtained through an analogous recursion. For this define the father coefficients $f_{j,k} = \boldsymbol{\mu}^t \boldsymbol{\varphi}_{j,k}$. Applying inner products to both sides of both equations in (9) gives the same relation between coefficients

$$\begin{aligned} \theta_{j,k} &= \tfrac{1}{\sqrt{2}} \left(f_{j+1,2k} - f_{j+1,2k+1} \right) \\ f_{j,k} &= \tfrac{1}{\sqrt{2}} \left(f_{j+1,2k} + f_{j+1,2k+1} \right). \end{aligned} \tag{10}$$

Using these equations iteratively over scales, starting from the values $f_{\log_2(n),k} = \mu_k$, for $k = 1, ..., n$ results in a fast and simple algorithm. As the FFT provides a big improvement over the naive implementation, this approach to Haar wavelet decomposition also reduces the $O(n^2)$ matrix multiplication illustrated in Section 2.1 of Chapter 1, to an $O(n)$ calculation, that is both slightly faster and also simpler than the FFT.

An important feature of this linear transformation is that it preserves power at each step in the sense that $\left(\theta_{j,k}^2 + f_{j,k}^2 \right) = \left(f_{j+1,2k}^2 + f_{j+1,2k+1}^2 \right)$. Allocation of signal power is the key to understanding the usefulness of wavelet transforms in general. In this Haar case, the transformation handles the power of "constant vectors" by putting it entirely into the father coefficient at the next lower frequency (this is an "averaging operation"), with a 0 contribution to the wavelet coefficient $\theta_{j,k}$ (this is a "differencing operation"). Hence the mother basis provides very effective compression of step functions (since there are many zero coefficients), and usually will be effective for estimating them.

The Haar basis is often not good at compressing signals that are smooth, as shown in the left panel of Figure 4, which uses the data from Figure 1. The result does not look much like the sum of two sinusoids, although the essential structure is recovered. Since the Haar basis is a set of step functions, signal compression is much better for the step function data from Figure 2, as shown in the left panel of Figure 4. Here the signal recovery is excellent (again the Universal Hard Threshold method was used to choose the set S of coefficients). Note that the behavior of the nonparametric regression estimate is rather different at the two jumps. The jump on the right is at the dyadic point $x = \frac{3}{4}$, and their is a very clean jump (because this feature is well described by a single Haar basis element). The jump on the left is at the very non-dyadic point $x = \frac{1}{3}$, which

involves a non-zero Haar coefficient at each scale. The error entailed by the estimation of all of these gives the slightly rougher effect near this jump.

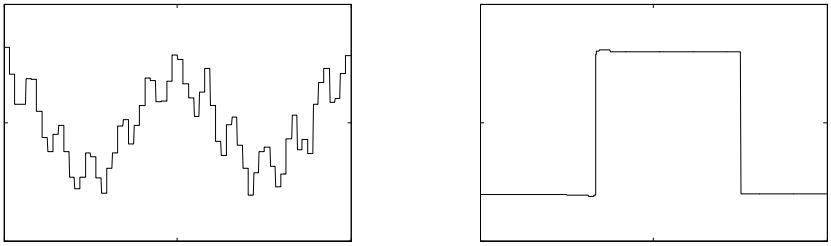

FIGURE 4. Haar basis nonparametric regression estimates, using hard thresholding, for the data in, Left panel: the left panel of Figure 1, Right Panel: the left panel of Figure 2.

Wavelet bases with "smoother" basis functions are much more useful for recovering smooth signals. These have a structure very similar to the Haar basis, in particular sharing the same indexing system, the "mother - father" relationships, the magnification property and analogous fast iterative algorithms for computation (with the same "pyramid structure"). Development of such bases has been fairly recent, perhaps because (in the continuous domain) orthonormal bases of compactly supported functions exist, but unfortunately have no closed form, and indeed seem to be smoothed versions of fractals, see Figure 1 of Chapter 1.

Generalization of the fast algorithm for the Haar basis to smoother basis functions follows from writing the relationships between coefficients of different scales in terms of operators:

$$\begin{array}{l} \boldsymbol{\theta}_j = (\downarrow 2_o)\,(\mathbf{hi} *_{F,F} \mathbf{f}_{j+1}) \\ \mathbf{f}_j = (\downarrow 2_e)\,(\mathbf{lo} *_{F,B} \mathbf{f}_{j+1}) \end{array},$$

where the coefficients of scale j have been combined into vectors as:

$$\boldsymbol{\theta}_j = \begin{pmatrix} \theta_{j,0} \\ \vdots \\ \theta_{j,2^j-1} \end{pmatrix}, \ \mathbf{f}_j = \begin{pmatrix} f_{j,0} \\ \vdots \\ f_{j,2^j-1} \end{pmatrix},$$

where $(\downarrow 2_e)$ and $(\downarrow 2_o)$ denotes the even and odd "decimation operator of order 2", which take "every other entry", i. e. for $m = 1, 2, ...$

$$(\downarrow 2_e)\begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_{2m-2} \\ x_{2m-1} \end{pmatrix} = \begin{pmatrix} x_0 \\ \vdots \\ x_{2m-2} \end{pmatrix}, \ (\downarrow 2_o)\begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_{2m-2} \\ x_{2m-1} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_{2m-1} \end{pmatrix},$$

where $*_{F,B}$ denotes a "back padded" circular filter operation

$$\begin{pmatrix} a_0 \\ \vdots \\ a_{\ell-1} \end{pmatrix} *_{F,B} \begin{pmatrix} b_0 \\ \vdots \\ b_{m-1} \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{\ell-1} a_i b_i \\ \sum_{i=0}^{\ell-1} a_i b_{1+i} \\ \vdots \\ a_0 b_n + \sum_{i=1}^{\ell-1} a_i b_{-1+i} \end{pmatrix},$$

and $*_{F,F}$ denotes a "front padded" circular filter operation

$$\begin{pmatrix} a_0 \\ \vdots \\ a_{\ell-1} \end{pmatrix} *_{F,F} \begin{pmatrix} b_0 \\ \vdots \\ b_{m-1} \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{\ell-2} a_i b_{m-\ell+1+i} + a_{\ell-1} b_0 \\ \sum_{i=0}^{\ell-3} a_i b_{m-\ell+2+i} + a_{\ell-2} b_0 + a_{\ell-1} b_1 \\ \vdots \\ \sum_{i=0}^{\ell-1} a_i b_{m-\ell+i} \end{pmatrix},$$

and where the "lo pass filter vector", which averages coefficients, thus passing the power of constants (i.e. the low frequency component of the sum of squares) into the $f_{j,k}$, is

$$\mathbf{lo} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and the "hi pass filter vector", which differences coefficients, thus passing the power of the residuals about the mean (i.e. the high frequency component of the sum of squares) into the $\theta_{j,k}$, is

$$\mathbf{hi} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

There are other ways to formulate these operations, but this particular form is chosen to make the coefficients correspond well to the location suggested by the indices.

Smoother wavelet bases are obtained by clever choices of the filter vectors **lo** and **hi**. The key to more smoothness is modification of the low pass filter to pass polynomials of a certain degree (the Haar filters are the degree 0 case). This entails filter vectors of length more than two, which entails more complicated orthogonality conditions. The shortest filter vectors which meet these requirements give the "Daubechies" family, see Table 6.1 of Daubechies (1992) , for numerical values of the coefficients of $\mathbf{lo} = (h_0, ..., h_{\ell-1})^t$ (where the symbol h is used for consistency with Daubechies (1992)). These basis functions are illustrated in Figure 1 of Chapter 1. Longer filter vectors allow bases that can satisfy additional requirements. For example, slightly longer filter vectors allow one degree of freedom that can be used to make the resulting basis functions "as symmetric as possible". This results in the family of filter vectors whose coefficients are in Table 6.3 of Daubechies (1992) , but the table values must be divided by $\sqrt{2}$ to give an orthonormal basis. The "Symmlet 8" ($N = 8$ in

Daubechies' notation) basis is shown in Figure 5, and its usefulness for non-parametric regression is demonstrated in Figure 6.

Only the low pass coefficients are given in Daubechies' tables, because the high pass coefficients, $\mathbf{hi} = (g_0, ..., g_{\ell-1})^t$ are simply related as:

$$g_i = (-1)^{\ell-1} h_{\ell-1-i}, \ \ i = 0, ..., \ell - 1$$

Insight into the inversion of the wavelet transform comes from realizing that it is just a matrix multiplication, by an orthonormal matrix (a Haar version of this matrix is shown in Section 2.1 of Chapter 1). Hence the inverse is multiplication by the transpose matrix. Plots of basis functions are easy to construct by calculating the inverse transform of unit vectors (i. e. vectors with a single 1, and all other entries 0, which are the transforms of the basis vectors).

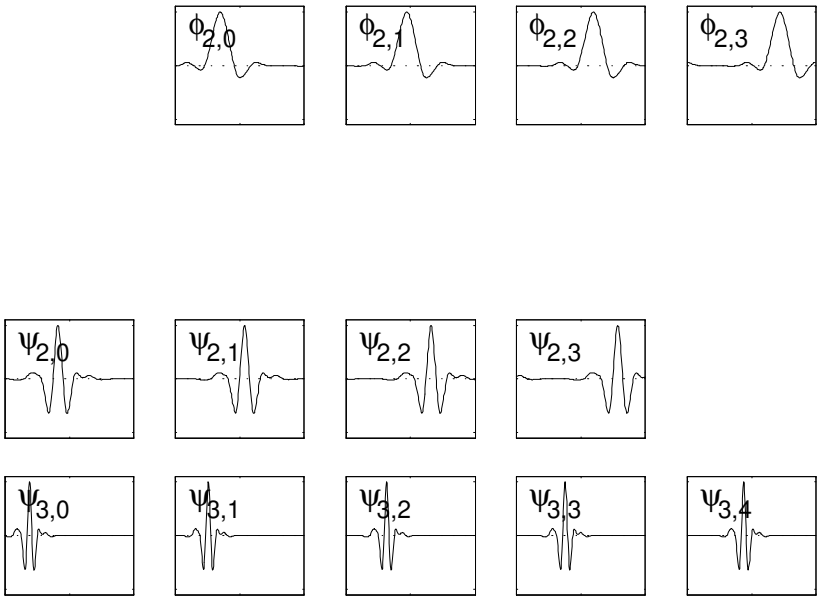

FIGURE 5: Symmlet 8 wavelet basis, with "father vectors" $\varphi_{2,0}$,...,$\varphi_{2,3}$ and "mother vectors" $\psi_{j,k}$, $j \geq 2$. As for the Haar basis, the row for scale $j = 3$ is incomplete, with panels for $k = 5, 6, 7$ truncated on the right side. The pattern continues below the bottom for scales $j = 4, 5, ...$

An organization of the Symmlet 8 basis is given in Figure 5. The layout is the same as Figure 3, except now instead of a single top level father basis element $\varphi_{0,0}$ the pyramid algorithm was stopped at level $j = 2$, resulting in the four father basis elements $\varphi_{2,0}, ..., \varphi_{2,3}$. Figure 5 suggests where the name "wavelet" came from. Note that these basis elements give the impression of being "small pieces of waves". But they have special properties, such as orthogonality and smoothness, that cannot be arrived at by simply trying to "slice up" the sinusoids that make up the Fourier basis.

An accessible introduction, from a different viewpoint, to more basic wavelet ideas can be found in Strang (1989) . A good source for deeper reading, including historical references, is Benedetto and Frazier (1994) . This book gives a good overview of continuous wavelet folklore in Section 1, and a readable presentation of broader and deeper aspects of discrete wavelets in Section 2. The following chapters consider many important variations of wavelet ideas.

2.2 Why Wavelets?

A common question is "why should one use the wavelet transform, when the Fourier transform has been such a useful workhorse?" The answer is that wavelet bases allow surprisingly good signal compression of a very wide range of signals. This point is illustrated in Figure 6, where the Symmlet 8 basis, shown in Figure 5, is used to denoise the same two data sets as used in the above figures.

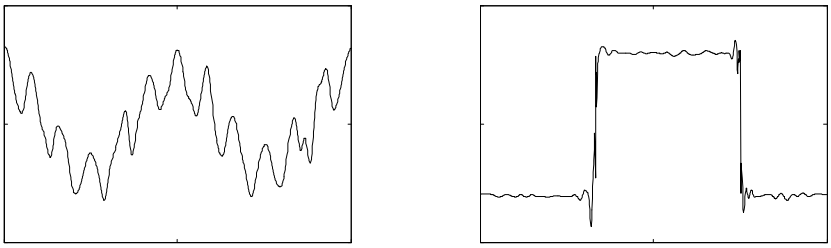

FIGURE 6. Symmlet 8 basis nonparametric regression estimates, using hard thresholding, for the data in, Left panel: the left panel of Figure 1, Right Panel: the left panel of Figure 2.

Note that using this basis, the signal recovery is excellent for *both* of these data sets. In contrast, the Fourier basis gave excellent recovery of the smooth sinusoidal target, but very poor performance for the step target, while the Haar basis gave excellent recovery of the step target, but poor recovery of the sinusoidal target. The key to understanding this is "signal compression". The Fourier basis represents the sinusoid with very few coefficients, but needs many for the step. The Haar represents the step with few coefficients, but needs many for the sinusoid. The power and elegance of smooth wavelet bases, such as the Symmlet 8 comes from being able to represent *both* types of signals in an efficient manner. This is the reason for the excitement that has surrounded the development of wavelet bases.

It is natural to expect that these impressive properties of wavelets come at some price. Perhaps the most serious downside to wavelets is that because they are flexible in both the time and frequency directions, they are rather sparse in each direction. For example, note that the number of "frequencies" represented is only $\log_2 n$ for the wavelet bases, while it is $n/2$ for the Fourier basis. The basis is also sparse in the time direction, as seen by noting that the peaks in

the $\psi_{2,k}$ row of Figure 5 appear in only four rather special locations. A signal with peaks that have a similar shape, but different location may not be so well compressed by this basis.

There are methods for addressing these, and other difficulties, although they all come at some cost. Many of these techniques, such as "wavelet frames", are discussed in Bennedetto and Frazier (1994). An approach to the location sparsity problem is the "nondecimated" wavelet transform (also called "stationary" and "shifted") discussed in Nason, Sapatinas and Sawczenko (1997) , which can be viewed as the set of all shifts of the wavelet basis, and is easily computed by eliminating the decimation operators $(\downarrow 2_e)$ and $(\downarrow 2_o)$ from the wavelet algorithm discussed above.

The already large literature on wavelet nonparametric regression is not surveyed here, but a very few suggested next references are Donoho and Johnstone (1995) and Donoho, Johnstone, Kerkyacharian and Picard (1995). See Marron, Adak, Johnstone, Neumann and Patil (1998) for an "exact risk" analysis of wavelet bases, which provides a different viewpoint on some of these ideas.

Acknowledgement 1 *The organization of the wavelet ideas used here mostly came from an informal presentation by David Donoho at the Oberwolfach meeting "Curves, Images and Massive Computation" in 1993. Sidney Resnick and Peter Mueller made a number of helpful comments.*

References

[1] Benedetto, J. J. and Frazier, M. W. (1994) *Wavelets: Mathematics and Applications*, CRC Press, Bocan Raton, Florida.

[2] Bloomfield, P. (1976) *Fourier analysis of time series, an introduction*, Wiley, New York.

[3] Daubechies, I. (1992) *Ten Lectures on Wavelets*,. SIAM, Philadelphia.

[4] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via Wavelet shrinkage, *J. Amer. Statist. Assoc.*, 90, 1200-1224.

[5] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia? *Jour. Roy. Statist. Soc. Ser. B*, 57, 301–369.

[6] Marron, J. S., Adak, S., Johnstone, I. M, Neumann, M. and Patil, P. (1998) Exact risk analysis of wavelet regression, *Journal of Computational and Graphical Statistics*,7, 278-309.

[7] Nason, G.P., Sapatinas, T. and Sawczenko, A. (1997) Statistical modelling of time series using non-decimated wavelet representations, Technical Report, University of Bristol.

[8] Strang, G. (1989) Wavelets and dilation equations: a brief introduction, *SIAM Review*, 31, 614-627.